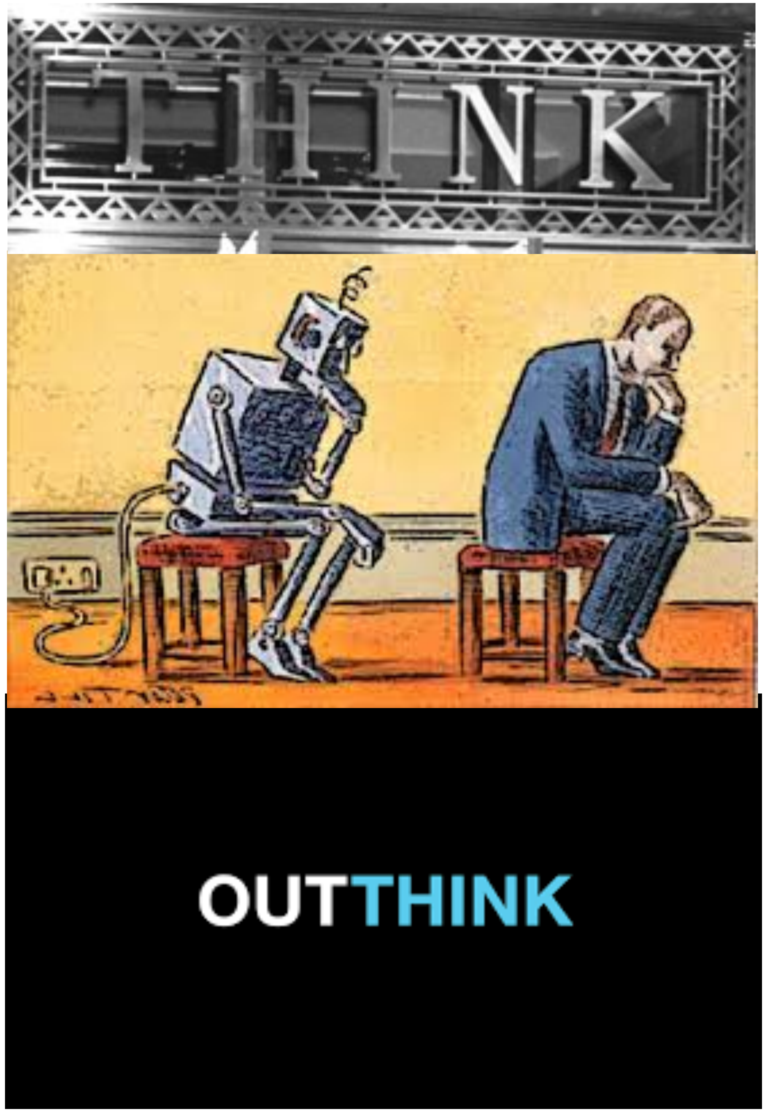
Knowledge Extraction, Representation and Reasoning Research Prospectus

From Linguistics to Knowledge

Vijay Saraswat

(DRAFT v 0.3)

January 2017



The Upton Vision

Upton: Achieving Professional Facility in Deep Domains

"[Edison] was brimming over with ideas but needed someone with advanced mathematical skills who could do calculations and research the scientific literature to help solve intractable problems. Despite his inveterate suspicion of academic scientists, Edison found Upton highly engaging and quite useful."

Professional facility: Operate as an assistant with the mastery of professionals in the field

"..Francis Robbins Upton, the very first student to officially earn, by examination, a graduate degree from Princeton. He received a Masters of Science in 1877."

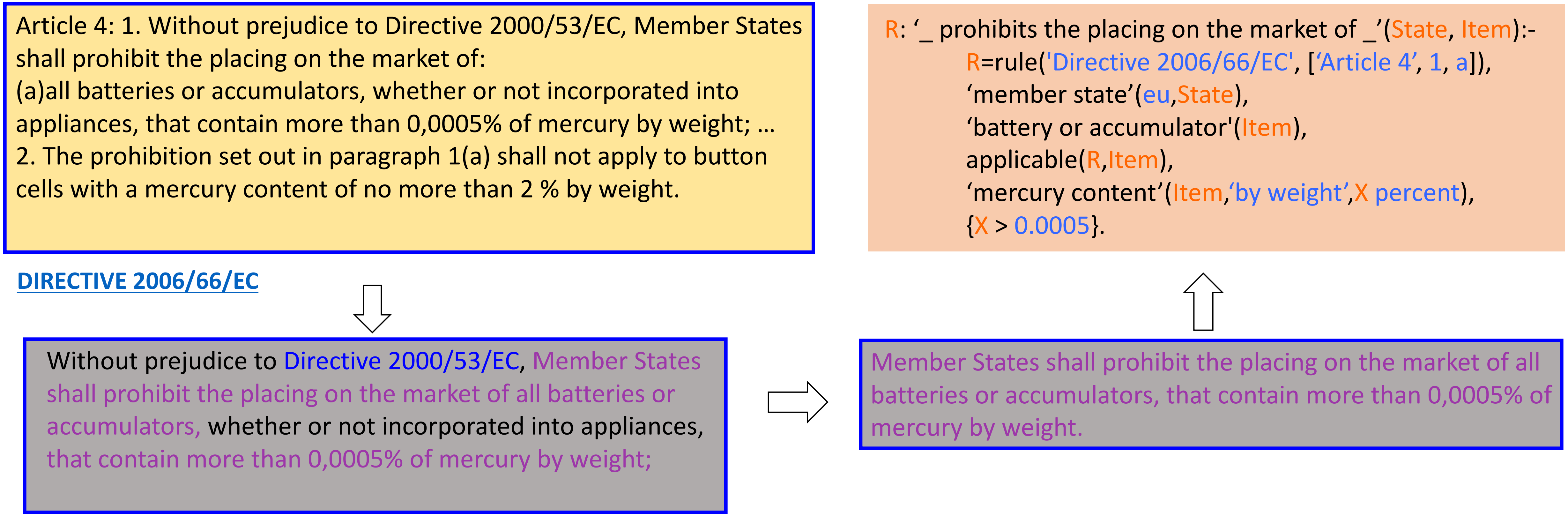
Deep domains: Domains with significant amount of pre-existing (formalizable) knowledge

Key Principles

- Support continuous ingestion of documents
  - System always has latest data – attractive to users.
- Early on, get a usable (cloud-connected) system in the hands of real users, designed for continuous feedback.
  - Users are legal, compliance, financial professionals – not linguists / computer scientists or KR engineers.
  - Build system around key functionality critical for users
  - Support teams – appropriate work-flow
  - Ensure user can correct / provide feedback for every system response, in terms that make sense to them.
  - Do this early to maximize learnings – putting a-man-in-the-box initially, if necessary.
- Design system **to learn continuously**
- **Support** automated conversion from input text to a structured form (which is deterministically convertible into logical form)

Support short-term projects – in context of over-arching research thrust

Example Task: Professional Question Answering



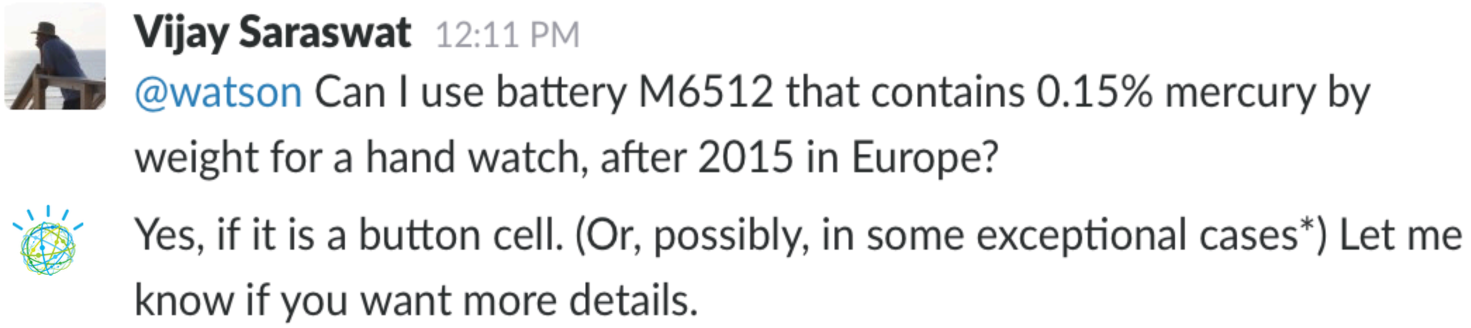
Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:
(a)all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight; …
2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight.

DIRECTIVE 2006/66/EC

Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

Member States shall prohibit the placing on the market of all batteries or accumulators, that contain more than 0,0005% of mercury by weight.

```
R: ‘_ prohibits the placing on the market of _’(State, Item):-
        R=rule('Directive 2006/66/EC', [‘Article 4’, 1, a]),
        ‘member state’(eu,State),
        ‘battery or accumulator'(Item),
        applicable(R,Item),
        ‘mercury content’(Item,‘by weight’,X percent),
        {X > 0.0005}.
```



Vijay Saraswat 12:11 PM

@watson Can I use battery M6512 that contains 0.15% mercury by weight for a hand watch, after 2015 in Europe?

Yes, if it is a button cell. (Or, possibly, in some exceptional cases*) Let me know if you want more details.

Actually, it is more complex!

DIRECTIVE 2013/56/EU

Article 1
Directive 2006/66/EC is amended as follows:
(1) Article 4 is amended as follows:
(a) paragraph 2 is replaced by the following:
’2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight until 1 October 2015.’;
…

Subsequent regulations invalidate some portions of old regulation.

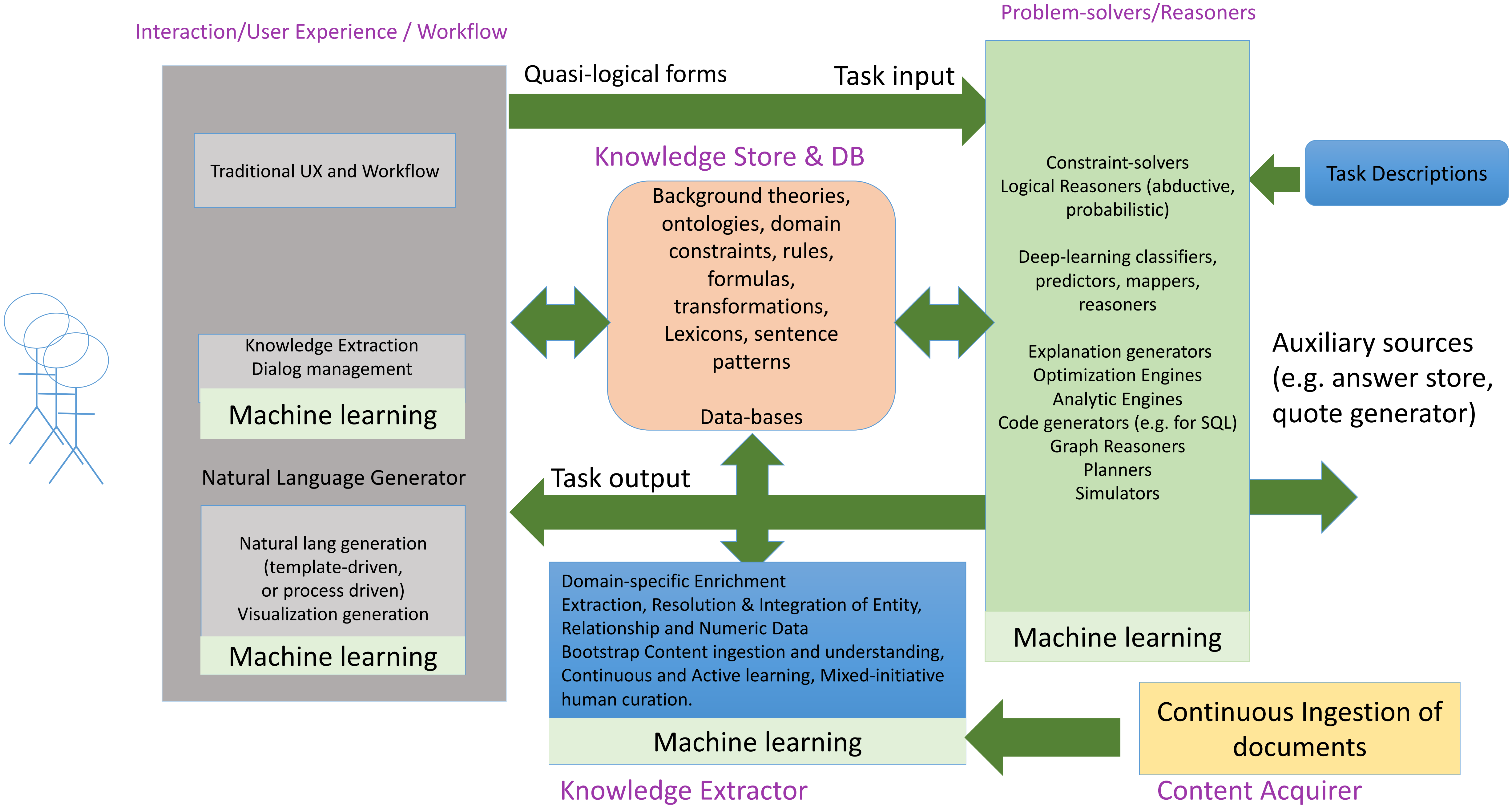
Regulation does this by simply replacing clauses in old regulations with new clauses.

Statistical techniques do not understand “quotes”!
Logical techniques can.

Task: Construct compositionally* the currently active obligations

* Compositionality: Conjunction of obligations extracted separately from Reg A and Reg B should give the right result even if Reg B rupdates Reg A.

Task-based reasoning architecture



Interaction/User Experience / Workflow

Traditional UX and Workflow

Knowledge Extraction
Dialog management

Machine learning

Natural Language Generator

Natural lang generation
(template-driven,
or process driven)
Visualization generation

Machine learning

Quasi-logical forms

Task input

Knowledge Store & DB

Background theories, ontologies, domain constraints, rules, formulas, transformations, Lexicons, sentence patterns

Data-bases

Task output

Problem-solvers/Reasoners

Constraint-solvers
Logical Reasoners (abductive, probabilistic)

Deep-learning classifiers, predictors, mappers, reasoners

Explanation generators
Optimization Engines
Analytic Engines
Code generators (e.g. for SQL)
Graph Reasoners
Planners
Simulators

Machine learning

Task Descriptions

Auxiliary sources (e.g. answer store, quote generator)

Domain-specific Enrichment
Extraction, Resolution & Integration of Entity, Relationship and Numeric Data
Bootstrap Content ingestion and understanding,
Continuous and Active learning, Mixed-initiative human curation.

Machine learning

Knowledge Extractor

Continuous Ingestion of documents

Content Acquirer

User Experience

Shared “AI Inside” Workbench for real users doing real work

Show related documents

Extract Obligation

Show Indirect obligations

Compare Obligation

Merge Obligation

Find relevant obligations

Correct annotation

Map to my ontology

Generate current obligations

Ask me!

Search

Topic: MiFiD2, Investment Firms

7. All multilateral systems in financial instruments shall operate either in accordance with the provisions of Title II concerning MTFs or OTFs or the provisions of Title III concerning regulated markets.

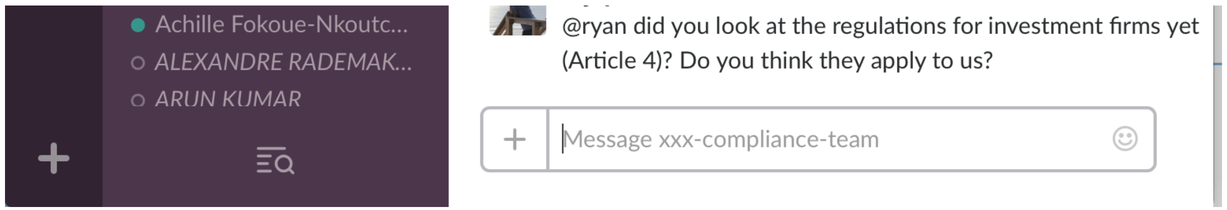
Any investment firms which, on an organised, frequent, systematic and substantial basis, deal on own account when executing client orders outside a regulated market, an MTF or an OTF shall operate in accordance with Title III of Regulation (EU) No 600/2014.

ok

Any investment firm shall operate in accordance with Title III of Regulation (EU) No 600/2014 if

(a) They deal on own account when executing client orders outside a regulated market, an MTF or an OTF

(b) They do this on an organised, frequent, systematic and substantial basis



● Achille Fokoue-Nkoutc...

○ ALEXANDRE RADEMAK...

○ ARUN KUMAR

@ryan did you look at the regulations for investment firms yet (Article 4)? Do you think they apply to us?

Message xxx-compliance-team

The “new collar” workbench – Keystone project

From Linguistics to Knowledge

Linguistic phenomena in web conversations

how much would my quarterly payment be in total
so how much is the total quarterly payment ?
how much are the 5 quarterly payments
how much for buying the phone up front?
How much is the quarterly installment fee?

Short sentences with lots of expletives (noise words)

Long-distance dependencies: wh-fronting

Modification

Coordination – sentential, NP-, VP-, involving unsaturated predicates.

Anaphora

…

Moderately Complex – good initial test-bench

Linguistic phenomena in contracts

> The effective date of this order will be the later of the Effective Date noted herein or the date User Id is issued .
>
> Customer shall not share its user ID 's and passwords outside the U.S. , nor may it share information accessed under this Order with persons located outside the U.S .
>
> Customers may only make Services under this Order available to entities located in the United States that are subsidiaries , divisions or affiliates , wholly-owned or controlled by Customer (`` US Affiliates ’’) and identified on a `` Schedule of Affiliates '' attached to this Order and that are not currently eligible to receive any Services included herein under an existing agreement with XXX to support their respective US businesses .

Long-distance dependencies: relative clauses

- that are not currently eligible to receive any Services included herein under an existing agreement with XXX to support their respective US businesses

Complex (vs Simple / Compound) sentences – typically less than ten clauses.

Modification - effective date

Coordination –

- user ID’s and passwords
- the later of the Effective Date noted herein or the date User Id is issued
- not share its user ID 's and passwords outside the U.S. , nor may it share information accessed under this Order with persons located outside the U.S

Anaphora -- its user ID’s and passwords

Moderately Complex – good initial test-bench

…

Linguistic phenomena in regulations

> v) Domestic scheduled commercial banks (other than RRBs) are permitted to open branches, Administrative offices, Central Processing Centres (CPCs) and Service branches in Tier 2 to Tier 6 centres (with population up to 99,999 as per Census 2001 - details of classification of centres tier-wise furnished in Annex 5) and in rural, semi-urban and urban centres in North Eastern States and Sikkim, and to open mobile branches in Tier 3 to Tier 6 centres (with population up to 49,999 as per Census 2001) and in rural, semi-urban and urban centres in North Eastern States and Sikkim without permission from Reserve Bank of India in each case, subject to reporting.

T*arget formalization – currently manually generated*.

Single sentence may be split across multiple paragraphs, with multiply nested bullets

Complex (vs Simple / Compound) sentences – may have a large number of clauses

Modification

Coordination – sentential, NP-, VP-, involving unsaturated predicates

Anaphora -- which

Long-distance dependencies: e.g. relative clauses, wh-questions

Quantifiers

…

Complex and extremely challenging

Linguistic phenomena in regulations

Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:

(a) all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

(b) portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight.

Single sentence may be split across multiple paragraphs, with multiply nested bullets

Complex (vs Simple / Compound) sentences – may have a large number of clauses

Modification

Coordination – batteries or accumulators

Negation – whether or not

Anaphora -- which

Long-distance dependencies -- relative clauses:
whether or not incorporated into appliances
That contain more than … by weight

Quantifiers:
All batteries or accumulators…

Complex and extremely challenging

Linguistic phenomena in prospectuses

1) a) The Fund may exclusively invest in:

…

vii) Money market instruments other than those dealt in on a Regulated Market, if the issue or the issuer of such instruments are themselves regulated for the purpose of protecting investors and savings, and provided that such instruments are:

a.issued or guaranteed by a central, regional or local authority or by a central bank of an EU Member State, the European Central Bank, the EU or the European Investment Bank, a non-EU Member State or, in case of a Federal State, by one of the members making up the federation, or by a public international body to which one or more EU Member States belong; or

b.issued by an undertaking, any securities of which are dealt in on Regulated Markets referred to in 1) a) i) and ii) above; or

c.issued or guaranteed by a credit institution subject to prudential supervision in accordance with criteria defined by European law or by a credit institution which is subject to and complies with prudential rules considered by the CSSF to be at least as stringent as those laid down by the European law; or

d.issued by other bodies belonging to the categories approved by the CSSF provided that investments in such instruments are subject to investor protection equivalent to that laid down in a. b. or c. above and provided that the issuer is a company whose capital and reserves amount to at least ten million Euro (EUR 10,000,000) and which presents and publishes its annual accounts in accordance with the fourth Directive 78/660/EEC, is an entity which, within a group of companies, is dedicated to the financing of the group or is an entity which is dedicated to the financing of securitisation vehicles which benefit from a banking liquidity line.

Single sentence may be split across multiple paragraphs, with multiply nested bullets

Complex (vs Simple / Compound) sentences – may have a large number of clauses

Modification

Coordination – sentential, NP-, VP-, involving unsaturated predicates

Anaphora -- which

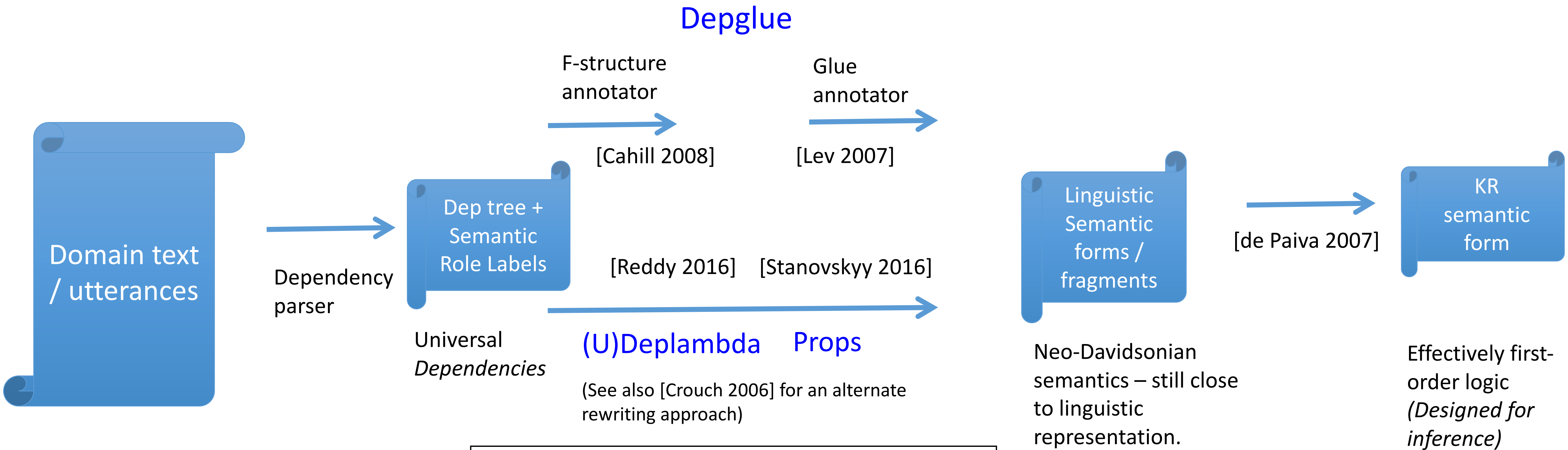
Long-distance dependencies: e.g. relative clauses, wh-questions

Quantifiers

…

Complex and extremely challenging

From linguistics to knowledge



Depglue

F-structure annotator

[Cahill 2008]

Glue annotator

[Lev 2007]

Domain text / utterances

Dependency parser

Dep tree + Semantic Role Labels

Universal *Dependencies*

[Reddy 2016] [Stanovskyy 2016]

(U)Deplambda Props

(See also [Crouch 2006] for an alternate rewriting approach)

Linguistic Semantic forms / fragments

Neo-Davidsonian semantics – still close to linguistic representation.

[de Paiva 2007]

KR semantic form

Effectively first-order logic *(Designed for inference)*

Using unsupervised/supervised techniques, symbolic / neural techniques.
Note: Semantic Parsers typically require (utterance, semantic)* corpus. *Need to develop sembanks for our discourse domains*

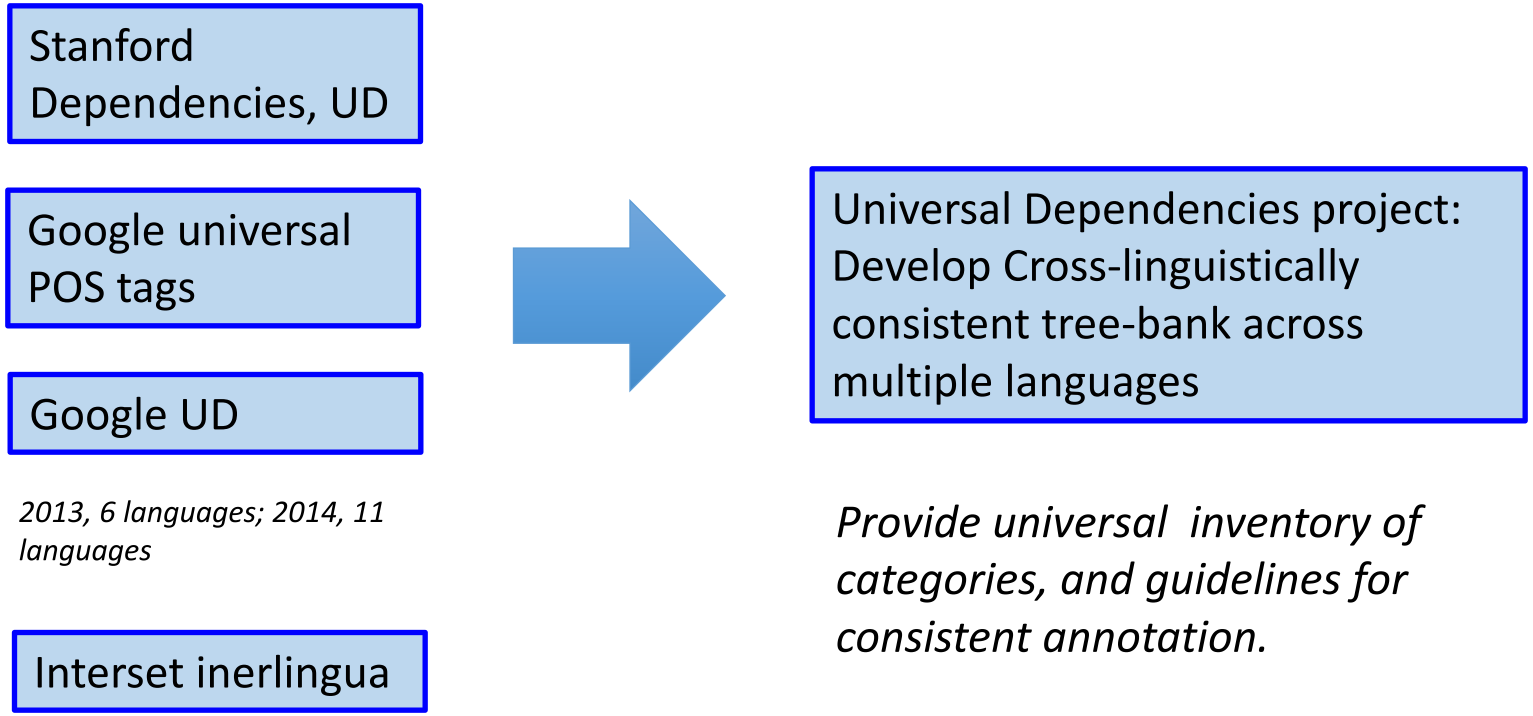
Issues to consider:

- Discourse (cf glue for DRT)
- Treatment of ambiguity (packed representation?)

Compliance | Legal | Financial | Contracts …

Note: PARC, Stanford researchers (Bobrow, de Paiva, Crouch, Karttunen, …) pioneered this line of attack (DARPA AQUAINT c 2007)

Dependency Parsers



Stanford Dependencies, UD

Google universal POS tags

Google UD

2013, 6 languages; 2014, 11 languages

Interset inerlingua

Universal Dependencies project: Develop Cross-linguistically consistent tree-bank across multiple languages

Provide universal inventory of categories, and guidelines for consistent annotation.

V1.2 has 37 treebanks in 33 languages

Core principle: “Meaning” of a sentence encoded in terms of relation between surface form tokens.

http://universaldependencies.org

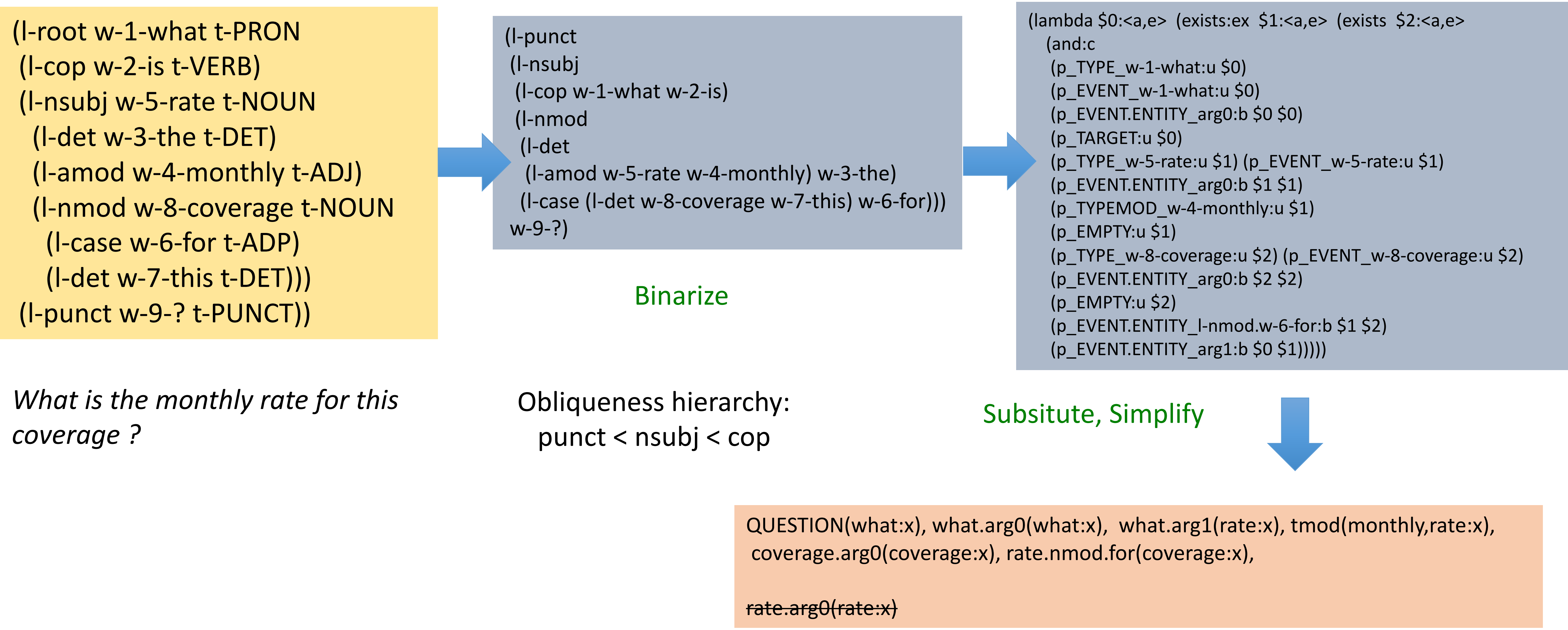
What is my monthly payment

```
(l-root w-1-what t-PRON
 (l-cop w-2-is t-VERB)
 (l-nsubj w-5-payment t-NOUN
   (l-nmod:poss w-3-my t-PRON)
   (l-amod w-4-monthly t-ADJ)))
```

What is the monthly rate for this coverage ?

```
(l-root w-1-what t-PRON
 (l-cop w-2-is t-VERB)
 (l-nsubj w-5-rate t-NOUN
   (l-det w-3-the t-DET)
   (l-amod w-4-monthly t-ADJ)
   (l-nmod w-8-coverage t-NOUN
     (l-case w-6-for t-ADP)
     (l-det w-7-this t-DET)))
 (l-punct w-9-? t-PUNCT))
```

UDepLambda [Reddy 2016; forthcoming]



```
(l-root w-1-what t-PRON
 (l-cop w-2-is t-VERB)
 (l-nsubj w-5-rate t-NOUN
   (l-det w-3-the t-DET)
   (l-amod w-4-monthly t-ADJ)
   (l-nmod w-8-coverage t-NOUN
     (l-case w-6-for t-ADP)
     (l-det w-7-this t-DET)))
 (l-punct w-9-? t-PUNCT))
```

What is the monthly rate for this coverage ?

```
(l-punct
 (l-nsubj
  (l-cop w-1-what w-2-is)
  (l-nmod
   (l-det
    (l-amod w-5-rate w-4-monthly) w-3-the)
   (l-case (l-det w-8-coverage w-7-this) w-6-for)))
 w-9-?)
```

Binarize

Obliqueness hierarchy:
punct < nsubj < cop

```
(lambda $0:<a,e>  (exists:ex  $1:<a,e>  (exists  $2:<a,e>
   (and:c
    (p_TYPE_w-1-what:u $0)
    (p_EVENT_w-1-what:u $0)
    (p_EVENT.ENTITY_arg0:b $0 $0)
    (p_TARGET:u $0)
    (p_TYPE_w-5-rate:u $1) (p_EVENT_w-5-rate:u $1)
    (p_EVENT.ENTITY_arg0:b $1 $1)
    (p_TYPEMOD_w-4-monthly:u $1)
    (p_EMPTY:u $1)
    (p_TYPE_w-8-coverage:u $2) (p_EVENT_w-8-coverage:u $2)
    (p_EVENT.ENTITY_arg0:b $2 $2)
    (p_EMPTY:u $2)
    (p_EVENT.ENTITY_l-nmod.w-6-for:b $1 $2)
    (p_EVENT.ENTITY_arg1:b $0 $1)))))
```

Subsitute, Simplify

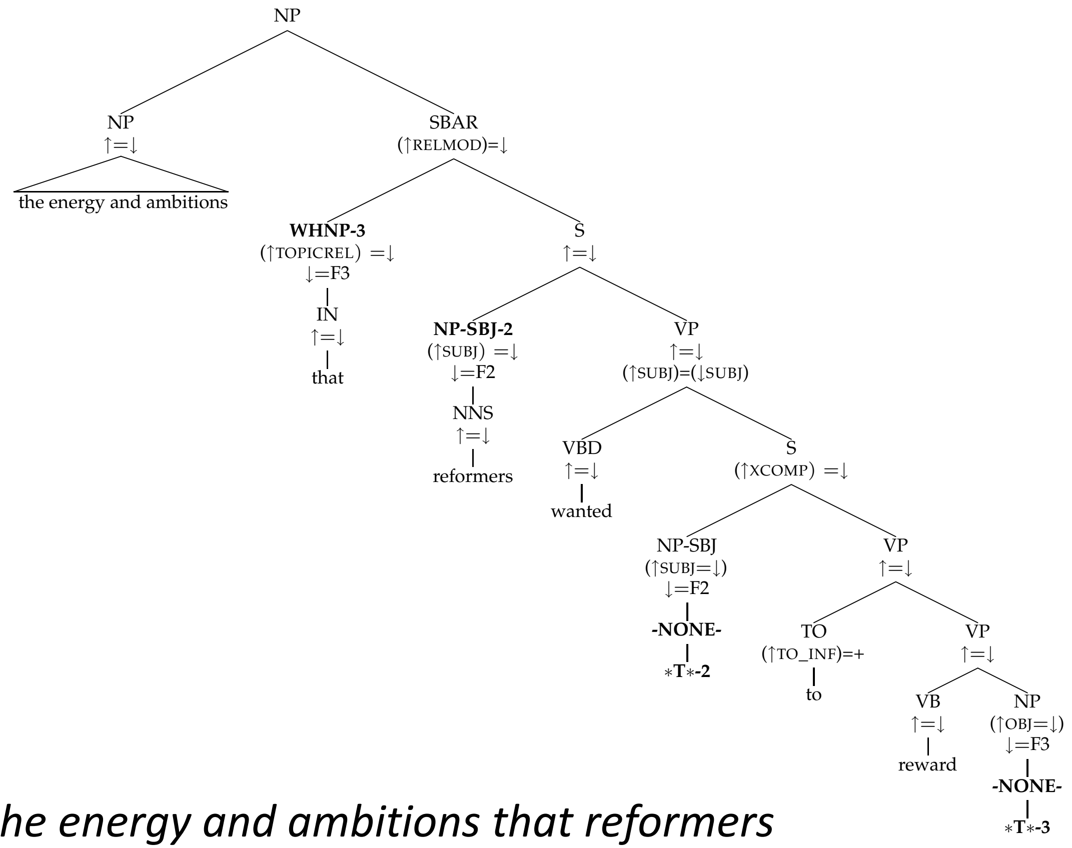
QUESTION(what:x), what.arg0(what:x), what.arg1(rate:x), tmod(monthly,rate:x), coverage.arg0(coverage:x), rate.nmod.for(coverage:x),

~~rate.arg0(rate:x)~~

Transform to application-friendly syntax

Generated using Stanford CoreNLP 3.7

What dependencies do not capture I/III



the energy and ambitions that reformers wanted to reward

root(ROOT-0, energy-2)
det(energy-2, the-1)
cc(energy-2, and-3)
conj:and(energy-2, ambition-4)
mark(wanted-7, that-5)
nsubj(wanted-7, reformers-6)
nsubj:xsubj(reward-9, reformers-6) // *T*-2
dep(energy-2, wanted-7)
mark(reward-9, to-8)
xcomp(wanted-7, reward-9)

Missing:
dobj:xsubj(reward-9, that-5) // *T*-3

Long Distance Dependencies

What dependencies do not capture II/III

```
(l-root w-2-wants t-VERB
 (l-nsubj w-1-everybody t-NOUN)
 (l-xcomp w-4-buy t-VERB
   (l-mark w-3-to t-PART)
   (l-dobj w-6-house t-NOUN
     (l-det w-5-a t-DET)))
 (l-punct w-7-. t-PUNCT))
```

Everybody wants to buy a house.

Adding **(l-xcomp w-4-buy t-VERB (l-nsubj w-1-everybody t-NOUN))** actually gives the meaning of *Everybody wants that everybody buys a house*

```
all(X, person(X),
    wants(X, a(Y,
      house(Y)&
         buys(X,Y))))
```

```
(l-root w-2-sleeps t-VERB
 (l-nsubj w-1-everybody t-NOUN)
 (l-cc w-3-or t-CONJ)
 (l-conj w-5-awake t-ADJ
   (l-cop w-4-is t-VERB))
 (l-punct w-6-. t-PUNCT))
```

Everybody sleeps or is awake.

Adding **(l-conj w-5-awake t-VERB (l-nsubj w-1-everybody t-NOUN))** actually gives the meaning of *Everybody sleeps or everybody is awake*

```
all(X, person(X),
   sleeps(X) | awake(X)))
```

Need to introduce variables (or some other linguistic device), i.e. move beyond core principle of dependency parsing.

Generalized Quantifiers

[Schuster & Manning 2016]

Quick aside: Quantifiers arise in regs!

Member States shall prohibit the placing on the market of all batteries or accumulators, that contain more than 0.005 percent of mercury by weight.

```
(l-root w-4-prohibit t-VERB
 (l-nsubj w-2-states t-PROPN
   (l-compound w-1-member t-PROPN))
 (l-aux w-3-shall t-AUX)
 (l-dobj w-6-placing t-VERB
   (l-det w-5-the t-DET)
   (l-nmod w-9-market t-NOUN
     (l-case w-7-on t-ADP)
     (l-det w-8-the t-DET)
     (l-nmod w-12-batteries t-NOUN
       (l-case w-10-of t-ADP)
       (l-det w-11-all t-DET)
       (l-cc w-13-or t-CONJ)
       (l-conj w-14-accumulators, t-NOUN)
       (l-acl:relcl w-16-contain t-VERB
         (l-nsubj w-15-that t-PRON)
         (l-dobj w-20-percent t-NOUN
           (l-nummod w-19-0.005 t-NUM
             (l-advmod w-17-more t-ADJ
               (l-mwe w-18-than t-ADP)))
           (l-nmod w-22-mercury t-NOUN
             (l-case w-21-of t-ADP)))
         (l-nmod w-24-weight. t-NOUN
           (l-case w-23-by t-ADP)))))))
```

```
R: '_ prohibits the placing on the market of _'(State, Item):-
        R=rule('Directive 2006/66/EC', ['Article 4', 1, a]),
        'member state'(eu,State),
        'battery or accumulator'(Item),
        applicable(R,Item),
        'mercury content'(Item,'by weight',X percent),
        {X > 0.0005}.
```

Generalized Quantifiers

What dependencies do not capture III/III

```
(l-root w-5-carrying t-VERB
 (l-nsubj w-1-sue t-PROPN
   (l-cc w-2-and t-CONJ)
   (l-conj w-3-mary t-PROPN))
 (l-aux w-4-are t-AUX)
 (l-dobj w-7-piano t-NOUN
   (l-det w-6-a t-DET))
 (l-punct w-8-. t-PUNCT))
```

```
a(X,  piano(X),
        carrying({sue,mary}, X))
```

```
a(X,  piano(X), carrying(sue, X)) &
a(X,  piano(X), carrying(mary, X))
```

Sue and Mary are carrying a piano.

Adding **(l-root w-5-carrying t-VERB (l-nsubj w-1-sue t-PROPN))**
actually gives the distributive interpretation:
Sue is carrying a piano and Mary is carrying a piano.

Need to support non-distributive interpretations of conjoined subjects

[Schuster & Manning 2016]

Conjoined subjects

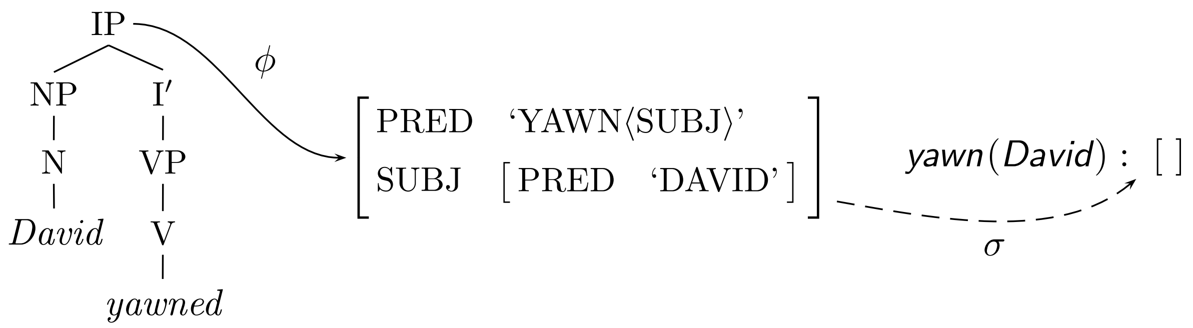
From Dependency Graphs to F-structures

- Starting in the late 70s, extensive work by linguists, logicians, computer scientists has led to deep computational theories of natural languages.
  - Typically, they address complex linguistic phenomena, e.g. long range dependencies.
  - Exemplars: LFG, GPSG, HPSG, CCG, …
- Generally, based on linguistics (non-transformational, cross-language) and logic (type theory, constraints), integrated now with statistical parsers.
- LFG, one of the premier such frameworks, has a well developed framework for semantic analysis (features + glue) that already addresses some of the conceptual difficulties with dependency graphs.
  - CCG is another influential framework
    - focuses on working directly with linear order of words in utterance

Core Proposal: Evolve dependency graphs to F-structures.

Deep Grammar Formalisms: LFG

(47) *David yawned.*



Constituent structure

.

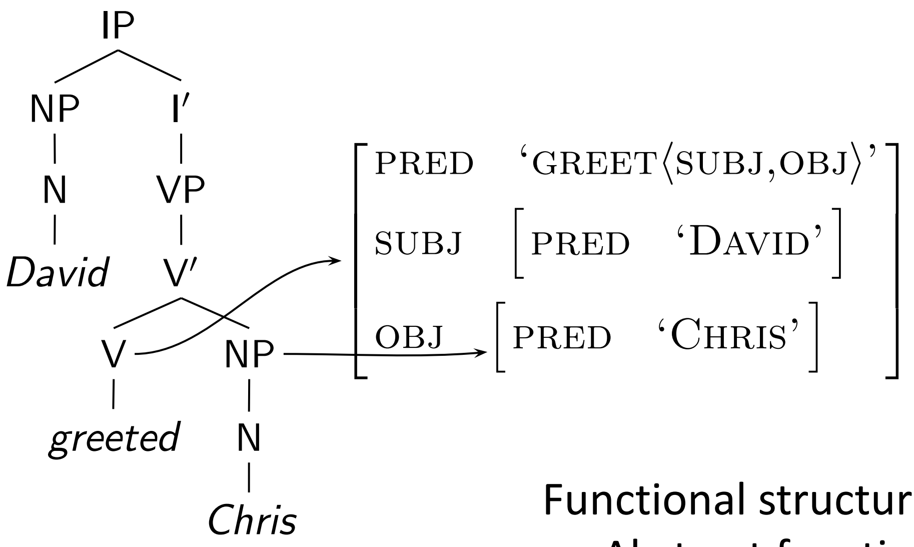
Functional structure

Semantic structure

- Deep grammars map strings to meaning representations
  - dependency structures
  - predicate-argument structure
  - simple logical forms
- Lexical-Functional Grammar (LFG): one of the oldest and most well-developed.
  - Organized around lexical and functional structure (not transformational, like Chomsky’s work).
  - Simultaneous levels of analysis – structural, functional, semantic, with projections and constraints tying them together

C-structures vs f-structures

LFG's c-structure and f-structure



Functional structure

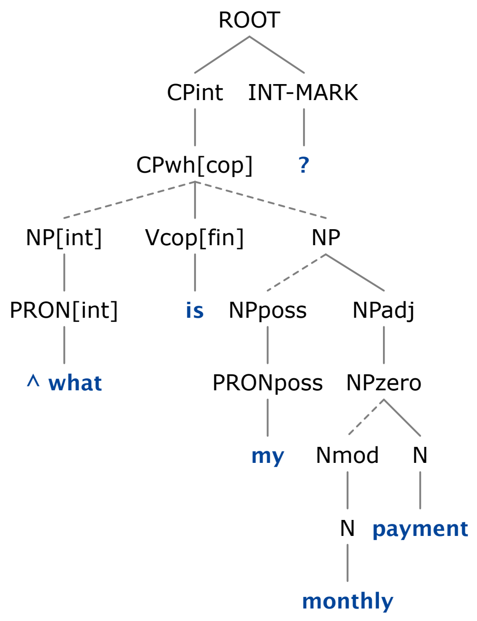
- Abstract functional syntactic organization of the sentence, familiar from traditional grammatical descriptions (Subject, Object, Adjunct),
- Representing syntactic predicate-argument structure and functional relations like subject and object.
- Theorized as cross-linguistically uniform.

Constituent structure is the overt, more concrete level of linear and hierarchical organization of words into phrases.

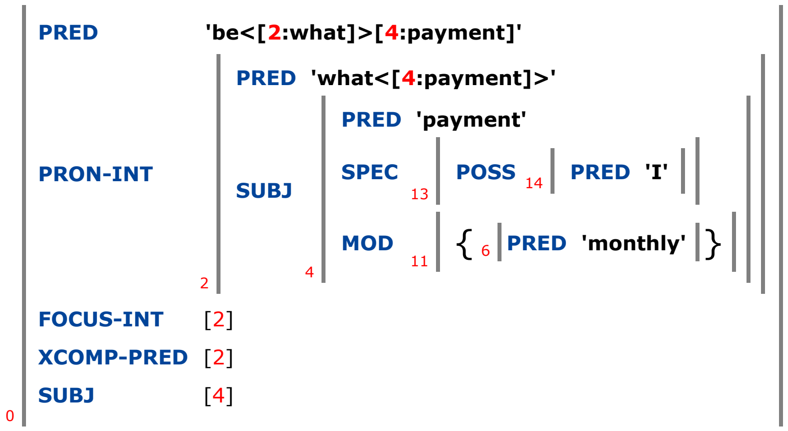
From [Dalrymple 2009]

C-structures vs f-structures

C-structure



F-structure



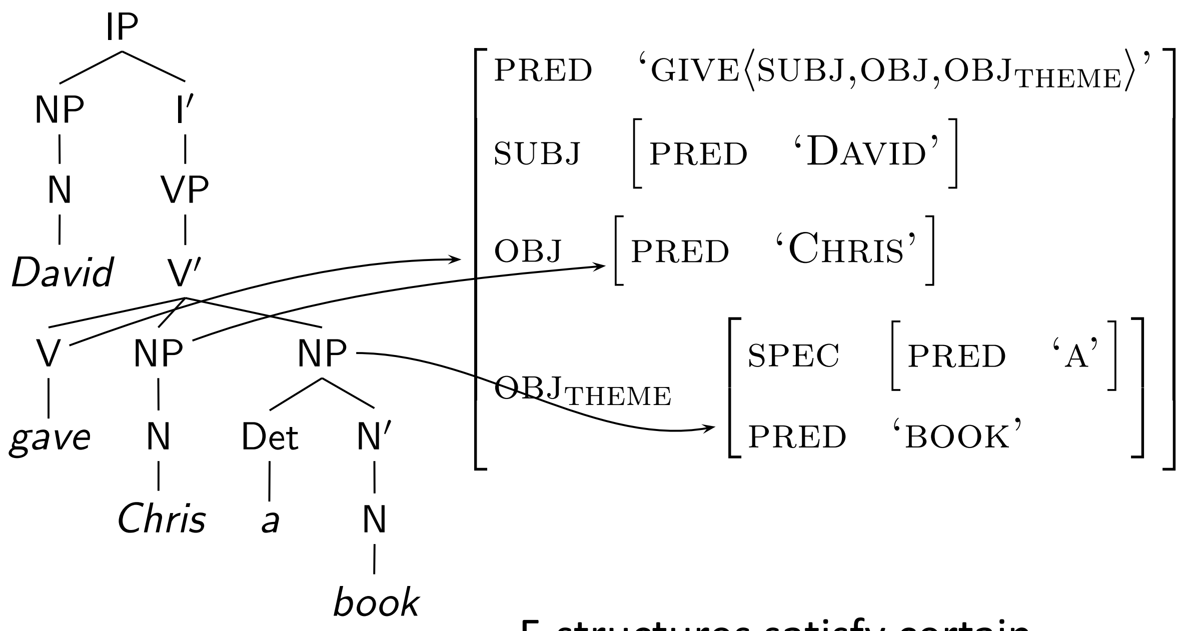
Constituent structure is the overt, more concrete level of linear and hierarchical organization of words into phrases.

- Abstract functional syntactic organization of the sentence, familiar from traditional grammatical descriptions (Subject, Object, Adjunct),
- Representing syntactic predicate-argument structure and functional relations like subject and object.
- Theorized as cross-linguistically uniform.

From [Dalrymple 2009]

F-structures

Complements of Lexical Categories



F-structures satisfy certain semantic constraints: Coherence, Completeness, Consistency.

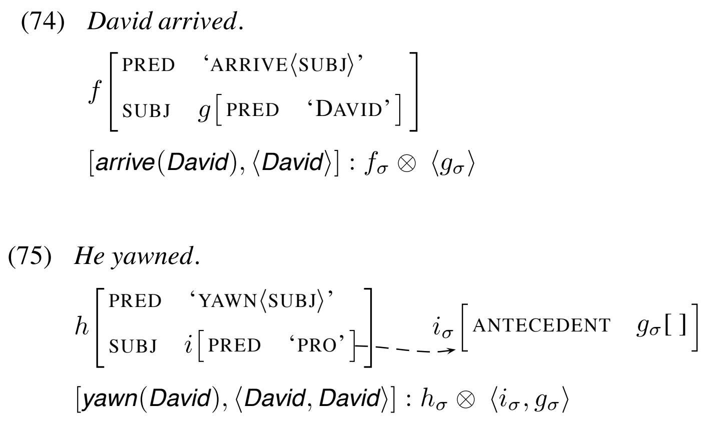
From [Dalrymple 2009].
See [Dalrymple 2001] for a thorough accout

- F-structure reflects the grammatical structure of the sentence (independently of, but coordinated with) c-structure.
- LFG assumes a universally available inventory of grammatical functions:
- SUBJ, OBJ, OBJ$_\theta$, COMP, XCOMP, OBL$_\theta$, ADJ, XADJ
- Θ: semantic roles, such as THEME, SOURCE, GOAL

- F-structures are nested, reentrant attribute value matrices. Values may be sets (with attributes). A very rich vocabulary of descriptions of F-structures has been developed over 30+ years: *equations, disjunctions, negations, optional constraints, (negative) existential constraints , (inside out) functional uncertainty, set descriptions, PCASE ("eval"), non-distributive features*

- Very rich linguistic phenomena modeled using F-structures

From f-structures to logical forms using glue



(74) *David arrived.*

$$f\begin{bmatrix}\text{PRED} & \text{`ARRIVE}\langle\text{SUBJ}\rangle\text{'}\\ \text{SUBJ} & g\begin{bmatrix}\text{PRED} & \text{`DAVID'}\end{bmatrix}\end{bmatrix}$$

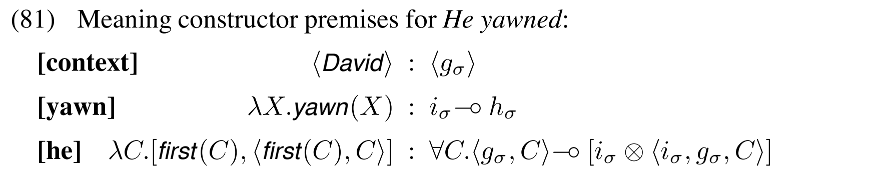
$[\mathit{arrive}(\mathit{David}), \langle \mathit{David}\rangle] : f_\sigma \otimes \langle g_\sigma\rangle$

(75) *He yawned.*

$$h\begin{bmatrix}\text{PRED} & \text{`YAWN}\langle\text{SUBJ}\rangle\text{'}\\ \text{SUBJ} & i\begin{bmatrix}\text{PRED} & \text{`PRO'}\end{bmatrix}\end{bmatrix} \dashrightarrow i_\sigma\begin{bmatrix}\text{ANTECEDENT} & g_\sigma[\;]\end{bmatrix}$$

$[\mathit{yawn}(\mathit{David}), \langle \mathit{David}, \mathit{David}\rangle] : h_\sigma \otimes \langle i_\sigma, g_\sigma\rangle$

Proof:

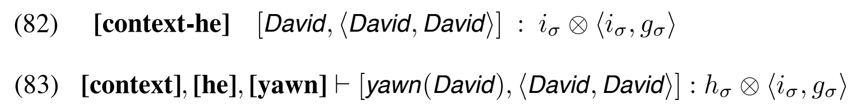


(81) Meaning constructor premises for *He yawned*:

[context] $\langle \mathit{David}\rangle : \langle g_\sigma\rangle$

[yawn] $\lambda X.\mathit{yawn}(X) : i_\sigma \multimap h_\sigma$

[he] $\lambda C.[\mathit{first}(C), \langle \mathit{first}(C), C\rangle] : \forall C.\langle g_\sigma, C\rangle \multimap [i_\sigma \otimes \langle i_\sigma, g_\sigma, C\rangle]$



(82) **[context-he]** $[\mathit{David}, \langle \mathit{David}, \mathit{David}\rangle] : i_\sigma \otimes \langle i_\sigma, g_\sigma\rangle$

(83) **[context]**, **[he]**, **[yawn]** $\vdash [\mathit{yawn}(\mathit{David}), \langle \mathit{David}, \mathit{David}\rangle] : h_\sigma \otimes \langle i_\sigma, g_\sigma\rangle$

From [Dalrymple 2001]

F-structure provides predicate/arg structure, but not scoped logical forms (w variables, quantifiers etc).

Glue offers a powerful compositional framework for meaning assembly, using deduction in linear logic.

Sematic contributions of components (lambda forms typed with propositional linear logic) are assembled into a term of a given type, via deduction. All terms that can be so constructed represent possible meanings for the utterance.

Glue is agnostic to the actual logic of meanings – one could use Montague's intensional logic, or some other application-dependent logic.

Glue has been shown to be remarkably powerful, handling wide range of semantic phenomena, see [Dalrymple 2001] … , quantification, intensional verbs, modification, coordination, anaphora, ...

Projects

Projects

More details on Linguistics To Knowledge Research

Deep Domain Parser Research

- Nature of regulatory text is very different from news cf [Morgenstern 2014].
    - Deontic irrealis mood vs realis mood
    - Far fewer mentions of named entities – and the entities are different (e.g. regulatory agencies and their organs, acts).
    - The meaning of a sentence often relates to text in other parts of the document (e.g. via references, use of definitions).
    - Significant use of abbreviations, references, scoped definitions.
    - Text is much more complex – very low Flesch scores, long complex sentences (many clauses)
    - Text may be structured – single sentence spread across paragraphs, bulleted lists.

- Fortunately, for the most part sentences are dry, precise, declarative and factual or deontic.
    - Text is intended to be clear and descriptive; ambiguity, if present, is deliberate
    - Little, if any, use of metaphors, similes, irony, allusion, sarcasm, satire, alliteration... ➔ we are not dealing with literature (whew!)

Cf [Lev 2007] structural semantics

Parser research

- Need strategy for dealing with long sentences, spread across multiple paragraphs. Look for techniques to shorten (cf improve Flesch readability metric) e.g.
  - Recognize and mask certain kinds of compound NPs e.g. Tier 2 to Tier 6, Domestic scheduled commercial banks (other than RRBs), branches / Central Processing Centres (CPCs) / Service branches
  - Deal with parenthetical remarks e.g., Tier 1 centres (centres with population of 1,00,000 and above as per 2001 Census)
  - Break up sentences – Sentence expansion.

1) Each depository institution which has a home office or branch office located within a primary metropolitan statistical area, metropolitan statistical area, or consolidated metropolitan statistical area that is not comprised of designated primary metropolitan statistical areas, as defined by the Department of Commerce shall compile and make available, in accordance with regulations of the Bureau, to the public for inspection and copying at the home office, and at least one branch office within each primary metropolitan statistical area, metropolitan statistical area, or consolidated metropolitan statistical area that is not comprised of designated primary metropolitan statistical areas, in which the depository institution has an office the number and total dollar amount of mortgage loans which were (A) originated (or for which the institution received completed applications), or (B) purchased by that institution during each fiscal year (beginning with the last full fiscal year of that institution which immediately preceded the effective date of this title).

#Sentences=1 #words=154 #syllables=329 Flesch score=-126.9

1) We define the designated areas for a depository institution as a primary metropolitan statistical area, metropolitan statistical area, or consolidated metropolitan statistical area that is not comprised of designated primary metropolitan statistical areas, as defined by the Department of Commerce.

Depository institutions with a home office or branch office located in designated areas shall compile and make available per conditions in (1.a) **the number and total dollar amount of mortgage loans handled by that institution** (as defined in 1.b) **during each fiscal year.** This obligation begins with the last full fiscal year of that institution which immediately preceded the effective date of this title.

(1.a) The required information shall be compiled and made available to the public for inspection and copying, in accordance with regulations of the Bureau, at the home office and at least one branch office within each designated area.

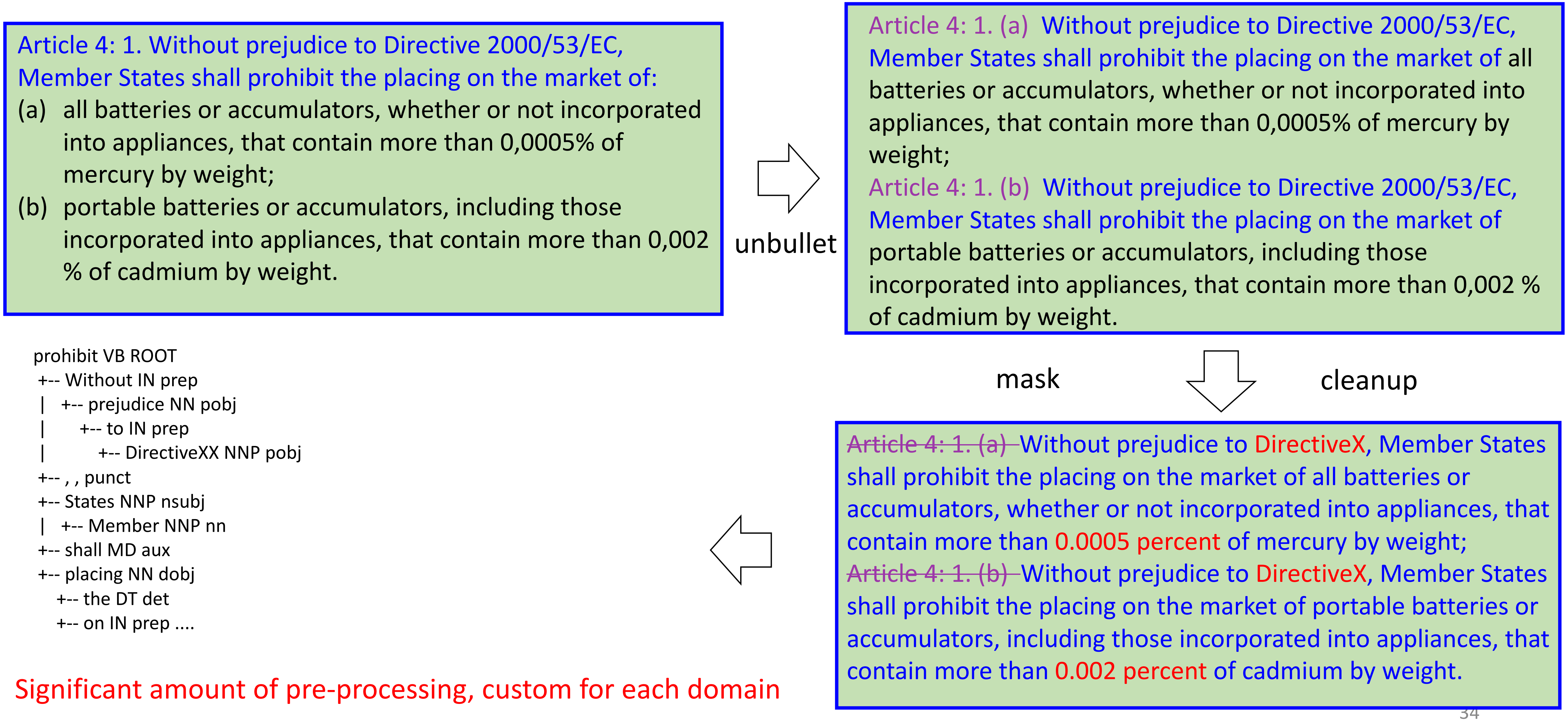
(1.b) The mortgage loans handled by an institution in a fiscal year are defined as loans which were (A) originated (or for which the institution received completed applications), or (B) purchased by that institution during that year.

#Sentences=9 #words=179 #syllables=349 Flesch score=30.2

Parser research

- Need to customize for our (professional) domains – e.g. handle
  - Different abbreviation styles
  - Quote marks e.g. for "licence" for opening branches (Note: Indian / Japanese English)
  - Conventions for introducing abbreviations e.g. This Evaluation Agreement (“Agreement”)
  - Material in bulleted lists
  - Nested definitions
  - Different citation styles e.g.
    - paragraph 1 (a) (i),
    - Directive 2000/53/EC,
    - referred to in 1) a) i) and ii) above
  - (Inline) Statement and use of definitions e.g.
    - “XXX Confidential Information” means the XXX Samples, specifications, and supporting documentation of the XXX Samples.
    - The term “Confidential Information” shall be used when referring to either party’s or both parties’ Confidential Information, as appropriate.
    - … which were (A) originated (or for which the institution received completed applications), or (B) purchased by that institution …

Parsing Strategy



Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:

(a) all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

(b) portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

unbullet

Article 4: 1. (a) Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

Article 4: 1. (b) Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

mask

cleanup

~~Article 4: 1. (a)~~ Without prejudice to DirectiveX, Member States shall prohibit the placing on the market of all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0.0005 percent of mercury by weight;

~~Article 4: 1. (b)~~ Without prejudice to DirectiveX, Member States shall prohibit the placing on the market of portable batteries or accumulators, including those incorporated into appliances, that contain more than 0.002 percent of cadmium by weight.

```
prohibit VB ROOT
 +-- Without IN prep
 |   +-- prejudice NN pobj
 |       +-- to IN prep
 |           +-- DirectiveXX NNP pobj
 +-- , , punct
 +-- States NNP nsubj
 |   +-- Member NNP nn
 +-- shall MD aux
 +-- placing NN dobj
     +-- the DT det
     +-- on IN prep ....
```

Significant amount of pre-processing, custom for each domain

Open-IE domain knowledge extraction

- (Dependency parsers already provide significant support for open information extraction. Build on it.)
- (Need to deal with structure in the domain – bulleted lists, scoped definitions etc.)
- (Different notion of named entities in these domains.)

Projects

Knowledge Representation and Reasoning

Deep Domain Knowledge Representation: Directed KR

- Regulatory, financial text typically has clear, definite logical content, typically about a specific relation
- Conditions typically refer to predicates defined in a background *domain theory* (i.e. branch/1, population/3, …), and use(numerical) constraints. May be complex (nested disjunctions, conjunctions).
- Directed KR: Questions are unanswerable by mapping into a known vocabulary (cf Rainy Day).
- For wide applicability, system should be able to use textual inference on (formally un-interpreted) text phrases.
- In the worst case, surface appropriate text directly to human. (Support human-in-the-loop mode.)

v) Domestic scheduled commercial banks (other than RRBs) are permitted to open branches, Administrative offices, Central Processing Centres (CPCs) and Service branches in Tier 2 to Tier 6 centres (with population up to 99,999 as per Census 2001 - details of classification of centres tier-wise furnished in Annex 5) and in rural, semi-urban and urban centres in North Eastern States and Sikkim, and to open mobile branches in Tier 3 to Tier 6 centres (with population up to 49,999 as per Census 2001) and in rural, semi-urban and urban centres in North Eastern States and Sikkim without permission from Reserve Bank of India in each case, subject to reporting.

```
rule([rbi2013, 3, v]):
'_ is permitted to open _ in _ with _ '(Bank, Branch, Loc, Condition) :-
  ('domestic scheduled commercial bank'(Bank), \+ ’RRB'(Bank)),
  (branch(Branch);
   'administrative office'(Branch);
   'CPC'(Branch);
   'service branch’(Branch)),
  (((tier(X, Loc), {X >= 2, X=< 6}),
   population(Loc, Pop, 'Census 2001'), {Pop =< 99999});
   (('rural centre'(Loc); 'semi-urban centre'(Loc); 'urban centre'(Loc)),
     state(Loc, State), (State='Sikkim'; 'North Eastern State'(State)))),
  Condition = {'no prior permission needed from RBI', 'subject to reporting'}.
```

<A bank> is permitted to open <a Branch> in <a Location> with <Conditions> if (certain predications hold)

Directed KR

- Regulatory, financial text typically has clear, definite logical content, typically about a specific relation
- KR framework must support ability to name rules (Article 4 1 a), and deny them, and deal with nested exceptions (not illustrated here).

Article 4: 1. Without prejudice to Directive 2000/53/EC, Member States shall prohibit the placing on the market of:

(a) all batteries or accumulators, whether or not incorporated into appliances, that contain more than 0,0005% of mercury by weight;

(b) portable batteries or accumulators, including those incorporated into appliances, that contain more than 0,002 % of cadmium by weight.

2. The prohibition set out in paragraph 1(a) shall not apply to button cells with a mercury content of no more than 2 % by weight.

```
R: '_ prohibits the placing on the market of _'(State, Item):-
        R=rule('Directive 2006/66/EC', ['Article 4', 1, a]),
        'member state'(eu,State),
        'battery or accumulator'(Item),
        applicable(R,Item),
        'mercury content'(Item,'by weight',X percent),
        {X > 0.0005}.
…
R: 'not applicable'(rule('Directive 2006/66/EC', ['Article 4', 1, a]), Item):-
          R=rule('Directive 2006/66/EC', ['Article 4', 2]),
        'button cell'(Item),
          applicable(R,Item),
        'mercury content'(Item, 'by weight', X percent),
        {X =< 2.0}.
```

<A State> prohibits the placing on the market of <an Item> if (certain predications hold)

Directed KR

- Regulatory, financial text typically has clear, definite logical content, typically about a specific relation.
- For wide applicability, system should be able to use textual inference on (formally un-interpreted) text phrases.
- In the worst case, surface appropriate text directly to human. (Support human-in-the-loop mode.)

General Investment Rules
1)a) The Fund may exclusively invest in:
1) a) iii) Recently issued transferable securities and money market instruments, provided that the terms of issue include an undertaking that application will be made for admission to official listing on a Regulated Market and such admission is secured within a year of the issue; and/or

```
‘_ may invest in _’(Fund, S) :-
        ('transferable security'(S); 'money market
instrument'(S)),
        'issue date'(S, D), reference_date(T), recent(D, T),
        'terms of issue'(S, Terms),
        includes(Terms, 'application will be made for
admission to official listing in a Regulated Market'),
        'application for admission to official listing in a
Regulated Market'(S, Event),
        date(Event, D), within_a_year(Date, D).
```

<A Fund> may invest in <a Security> if (certain predications hold)

Problem Solving (Longer term project)

- Attempt to deal with open-ended questions from user within the domain.
  - Need additional linguistic analysis / meaning generation to elaborate on user intent.
  - Need richer background theory.
  - Simplifying assumption: Key task is to work with domain terms – leave “open-ended common-sense reasoning” to humans.

- Assemble required theory for reasoning (from micro-theories in knowledge base).
- Construct response (with explanation)

Open-ended questions (within the given domain)

Explanation Generation

- Explanations need to be generated from proof-tree.
  - Ultimately grounded out in pieces of text from which the rules were generated + background theory.
- Explanation-generation is a critical NLG task
  - Needs to be done in the context of current user model (what is s/he interested in, what did they ask, what has already been explained to them…)
- Fundamentally, feedback should be acceptable in a fine-grained way, leading to improvement of the underlying system.

Timelines

(TBD, depending on level of effort)

Backup

Linguisic terminology

Linguistic terminology

- Modification – use of adjectives / adverbs to modify meanings.
  - E.g. Swedish man.
  - Note: modifiers can be recursive – e.g. apparently Swedish man.

- Coordination: frequently occurring complex syntactic struture that links together two or more elements (*conjuncts* or *conjoins*). Usually signalled by appearance of a *coordinator* (e.g. and, or, but). The totality of coordinator(s) and conjuncts in an instance of coordination is called a *coordination structure.*

Linguistic terminology

- A **relative clause** is a kind of subordinate clause that contains an element whose interpretation is provided by an antecedent on which the subordinate clause is grammatically dependent; that is, there is an anaphoric relation between the relativized element in the relative clause, and the antecedent on which it depends.
- In English, relative clauses are formed principally from relative pronouns. The basic pronouns are who (with derived form whom and whose), which and that.

Linguistic terminology

- **Anaphora**: use of an expression whose interpretation depends upon another expression in context (its antecedent or postcedent).
  - *I like it and so do they.*
  - *Units of UCITS authorised according to the UCITS Directive and/or other undertakings for collective investment ("UCI") within the meaning of the first and second indent of Article 1, paragraph (2) of the UCITS Directive, whether situated in an EU Member State or not, provided that […] such other UCIs have been authorised under laws which provide that they are subject to supervision considered by the CSSF to be equivalent to that laid down by European law and that cooperation between authorities is sufficiently ensured,*
- Ellipsis: omission of words from a clause that can be inferred from context.
  - *John can play the guitar and Mary$_{\textit{can play}}$ the violin.*

https://en.wikipedia.org/wiki/Linguistics

Linguistic terminology

- Long-distance dependencies
  - Topicalization: (Yodaisms – *Much to learn, you still have.*)
  - Wh-clauses: *What is your monthly payment? Which house is Bill living in?*
  - Relative clauses: *The man* ***who lives in this house*** *has not been seen for days.*

https://en.wikipedia.org/wiki/Linguistics

Penn Tree Bank POS tags

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb